Magyar nyelvű klinikai dokumentumok előfeldolgozása

Siklósi Borbála[1], Orosz György[1], Novák Attila[2]

[1] Pázmány Péter Katolikus Egyetem Információs Technológiai Kar, 1083 Budapest, Práter utca 50/a
e-mail: {siklosi.borbala, oroszgy}@itk.ppke.hu

[2] MorphoLogic Kft., 1116 Budapest, Kardhegy utca 5.
e-mail: novak@morphologic.hu

Kivonat A klinikai dokumentumok feldolgozásának első lépése azok strukturálása és normalizálása. Bemutatjuk, hogy a szerkezeti egységek hiányát hogyan tudtuk a formázási jegyek alapján automatikus transzformációkkal pótolni, illetve alapvető metainformációkat a folyó szövegből kinyerni. Ezután a korpusz szöveges részeit elválasztottuk a nem szöveges részektől, az így kapott halmazra automatikus helyesírás-javító, illetve javaslatgeneráló rendszert hoztunk létre. Módszerünk elsősorban a rendelkezésünkre álló korpusz statisztikai viselkedésére épül, de külső erőforrásokat is bevontunk a jobb minőség elérése végett. Az algoritmust két funkciója: a helyesírás-javítás, illetve a javaslatgenerálás alapján értékeltük ki. Beláttuk, hogy módszerünk a teljesen automatikus javításra pillanatnyilag önmagában nem alkalmas, azonban ez nem is volt cél, viszont minimális emberi közreműködéssel hatékonyan alkalmazható egy helyes orvosi-klinikai korpusz létrehozására.

Kulcsszavak: automatikus helyesírás-javítás, orvosi szövegfeldolgozás, szövegnormalizálás

1. Bevezetés

A legtöbb kórházban az orvosi feljegyzések tárolása csupán archiválás, illetve az egyes esetek dokumentálása céljából történik. Az így felhalmozódott adattömegek felhasználása jelenleg csupán az egyes betegek kórtörténetének visszakeresésére korlátozódik. A nyelvtechnológia, a számítógépes ontológiák és a statisztikai szövegfeldolgozó algoritmusok lehetővé tennék a folyó szövegekben rejlő összefüggések, rejtett struktúrák felfedését, a feljegyzésekben található információhalmaz elérését, abból tudás kinyerését.

Az angol nyelvterületen az ilyen irányú kutatások előrébb járnak, azonban alkalmazhatóságuk a magyar nyelv sajátosságai miatt sokszor nem egyértelmű, továbbá számos olyan nyelvi erőforrás, ami az angol nyelvre hozzáférhető, magyarra nem létezik. Az orvosi dokumentumok feldolgozása során nem csak a

magyar nyelv nyelvtani sajátosságait kell figyelembe venni, hanem az orvosi szövegekre különösen jellemző nehéz, olykor hiányos szintaktikai szerkezeteket, rövidítéseket, idegen kifejezéseket is kezelni kell.

Ezen tapasztalatok alapján fogalmazódott meg az igény, hogy a magyar nyelvű klinikai dokumentumok feldolgozását a más nyelveken már létező alkalmazások adaptálása, továbbfejlesztése és alkalmazhatóvá tétele révén aktívan kutatott területté tegyük, tekintettel a kutatás várható hasznára.

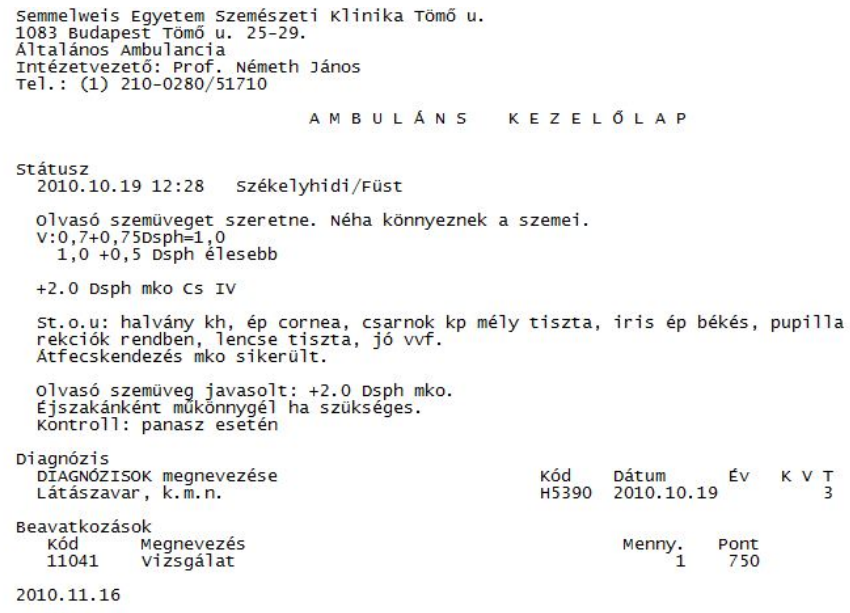
Hosszútávú célunk egy olyan keretrendszer készítése, amely orvosi dokumentumokat feldolgozva segíthet a klinikai szakembereknek új összefüggések feltárásában. Cikkünkben egy ilyen rendszer megvalósításának kezdeti lépéseit mutatjuk be. Az első probléma a rendelkezésünkre álló nyers orvosi szövegek egységes reprezentációjának kialakítása. Bár a meglévő klinikai dokumentumok láthatóan rendelkeznek struktúrával, de ezekre csak a formázás, illetve a tartalom értelmezése alapján lehet következtetni. Jelentős nehézség még a dokumentumokkal kapcsolatban, hogy készítőik nem fordítanak hangsúlyt a helyes és konzisztens fogalmazásra, tagolásra, helyesírásra. Így szükségesnek láttuk a dokumentumokban meglévő zaj (helyesírási hibák) csökkentését, ami akár orvosonként/asszisztensenként, illetve osztályonként is változó lehet.

Cikkünkben bemutatjuk a nyers orvosi dokumentumok feldolgozásakor alkalmazott algoritmusainkat, amelyekkel strukturális egységekre bontottuk a kórlapokat, és ezzel együtt a felszíni jegyekből könnyen meghatározható metainformációkat is kinyertünk, továbbá meghatároztuk az átfedő dokumentumrészeket. Ezek után bemutatjuk a szöveges és a nem szöveges részek elválasztására alkalmazott megoldásunkat, majd az automatikus helyesírás-javító rendszer első eredményeit ismertetjük.

2. A nyers dokumentumok strukturálása

Rendelkezésünkre állt a klinikai dokumentumok (kórlapok) egy rendezetlen halmaza. A szövegek struktúrájára csak a formázás, illetve a tartalom értelmezése alapján lehetett következtetni. Az alapvető tagoláson kívül – mely önmagában sem tekinthető egységesnek – nem voltak a további feldolgozás szempontjából használhatóan elkülönített egységek. Az adathalmaz jelentős része redundáns, az egyes esetek kórelőzményének minden korábbi fázisa a kórtörténet összes dokumentumában ismételten megjelenik, így a folyamat időben későbbi szakaszában készült leírások egyre hosszabbak, az összes előzmény másolása révén. Itt szintén tapasztalható volt az egységes rendszer hiánya, a folyamatok „összemásolása" többféle módon történt (időben korábbi/későbbi dokumentumok előrébb vagy hátrébb tolódása; diagnózisok elvetése/halmozása, stb.)

Mivel az eltérő szakterületek dokumentumainak felépítése eltérő, ezért elsőként a szemészeti dokumentumok feldolgozása indult el, melynek eredményei kisebb átdolgozással alkalmazhatóak lesznek más szakterületek, végül pedig általános orvosi szövegek feldolgozására.



```
Semmelweis Egyetem Szemészeti Klinika Tömő u.
1083 Budapest Tömő u. 25-29.
Általános Ambulancia
Intézetvezető: Prof. Németh János
Tel.: (1) 210-0280/51710

                          A M B U L Á N S   K E Z E L Ő L A P

Státusz
  2010.10.19 12:28   Székelyhidi/Füst

  Olvasó szemüveget szeretne. Néha könnyeznek a szemei.
  V:0,7+0,75Dsph=1,0
    1,0 +0,5 Dsph élesebb

  +2.0 Dsph mko Cs IV

  St.o.u: halvány kh, ép cornea, csarnok kp mély tiszta, iris ép békés, pupilla
  rekciók rendben, lencse tiszta, jó vvf.
  Átfecskendezés mko sikerült.

  Olvasó szemüveg javasolt: +2.0 Dsph mko.
  Éjszakánként műkönnygél ha szükséges.
  Kontroll: panasz esetén

Diagnózis
  DIAGNÓZISOK megnevezése                    Kód    Dátum      Év   K V T
  Látászavar, k.m.n.                         H5390  2010.10.19          3

Beavatkozások
   Kód      Megnevezés                                Menny.   Pont
   11041    Vizsgálat                                      1     750

2010.11.16
```

1. ábra. Egy eredeti dokumentum

2.1. XML-struktúra

A feldolgozás első lépéseként tehát szükséges volt a dokumentumok struktúrájának azonosítása és annak szabványos ábrázolása. Az egységek meghatározása egy egyszerű szabályalapú mintaillesztő eljárással történt, mely a rekordok szemmel is látható tagolására épül. Így a folyó szövegekben meglévő formázási elemeket transzformáltuk a szerkezetet meghatározó jellemzőkké. A kinyert struktúrák és metainformációk XML-struktúrában való tárolása során a dokumentumok felépítése a következőképpen alakult:

- Teljes eredeti: a teljes dokumentum szövegét eredeti formában is megtartottuk a későbbi megjelenítés egyszerűsítése céljából
- Tartalom: a dokumentumok szabad formájú szöveges részeit is tovább tagoltuk fejléc, diagnózisok, beavatkozások, javaslat, státusz, műtét, panasz, stb. részek megjelölésével.
- Metaadatok: a dokumentumok egyes részein alapvető automatikus módszerekkel jól felismerhető, a folyó szöveges részektől elkülönülő, adatokat tartalmazó egységeket nyertünk ki, ellátva őket az adatok típusára vonatkozó címkékkel. A következő metaadatokat nyertük ki: az adott dokumentum típusa (zárójelentés, kezelőlap stb); a dokumentumot kibocsátó osztály azonosítója; a táblázatos formában explicit módon megjelölt diagnózisok, illetve beavatkozások megnevezése és kódja.

– Egyszerű névelemek: a munkánk jelenlegi fázisában az egyszerű mintaillesztéssel kinyerhető névelemek (dátumok, orvosok, műtétek) megjelölése is megtörtént, azonban az erre alkalmazott módszerek finomítása és pontosítása még feltétlenül szükséges.
– Kórtörténet: az egyes betegek kórlefolyásának tárolása a klinikai adminisztrációs rendszer hiányosságai miatt jelenleg többféleképpen történik. Gyakori eset, hogy a kórelőzmény teljes szövege hozzáadódik az újabban keletkező dokumentumhoz, így folyamatosan egyre nagyobb dokumentumok kapcsolódnak egy pácienshez, melyek egymást tartalmazzák. Nincs egységes rendszer arra vonatkozóan sem, hogy a korábbi vizsgálatok leírása a dokumentumban előrébb vagy hátrébb – esetleg vegyesen – kerül be. Ennek ellenére megvalósult egy automatikus sorbarendezés, amelynek során minden dokumentumhoz eltároljuk az őt követő, és őt megelőző dokumentumokat – ha vannak ilyenek.

2.2. Szöveges részek elkülönítése

Az így kapott struktúra jól elkülöníti a dokumentumok egyes részeit, azonban korántsem elegendő ahhoz, hogy a szöveges részek önállóan kezelhetőek legyenek. Az általunk vizsgált szemészeti dokumentumokra különösen jellemzőek az esetek nagy részében túlnyomóan folyó szöveget tartalmazó szakaszokba ékelődő olyan nem folyó szöveg típusú részek, melyek az előfeldolgozás során zajként viselkednek. Ilyen részletek a laboreredmények, különböző számértékek, elválasztó karaktersorozatok, valamint csupán rövidítéseket, speciális jeleket tartalmazó megállapítások. Ezek kiszűrése szükséges volt ahhoz, hogy a nyelvi előfeldolgozás későbbi lépései során alkalmazott algoritmusok alapját képező korpusz előállítható legyen. Mivel azonban ezek a mintázatok önmagukban sem egységesek, különböző stílusú (feltételezhetően más-más orvos, illetve asszisztens szokásait tükröző) dokumentumok között még inkább változó módon szerepelnek, ezért szabályok, illetve mintafelismerés segítségével nem lehetett kiszűrni ezeket. A legkézenfekvőbb megoldásként klaszterezést alkalmaztunk. Mivel ezek a tartalmak sokrétűek, ezért mondatszegmentálást nem alkalmazhattunk, így a sorokra bontott dokumentumban kötöttük össze azokat, amik jó eséllyel egy egységet alkotnak. Ha egy sor nem mondatvégi írásjelre végződik, a rákövetkező sor pedig nem nagybetűvel és nem számmal kezdődik, illetve ha egy sor végén mondatközi írásjel van (vessző, pontosvessző), akkor a két sort összekötöttük.

Így megtartottuk azokat a mondattöredékeket, amik a felszíni jellemzőik alapján az elkülönítendő (nem szöveges) részekhez állnak közelebb. Az így megjelölt konkatenált sorokat K-means klaszterező algoritmussal csoportosítottuk. Célunk két diszjunkt halmaz létrehozása volt, de $k = 2$ esetén nem volt elég hatékony az elkülönítés. Mivel a jellemzőhalmaz módosításával nem sikerült célt érnünk, a klaszterek számának vizsgálata során optimális eredményt $k = 7$ esetén kaptunk, (A hét halmazból kettő tartalmazott szöveges részeket, a többi öt pedig különböző jellegű nem szöveges részeket) A klaszterezésnél használt jellemzőhalmaz, és az így létrejött tanítóanyag alkalmazásával a későbbiekben osztályozással is jól besorolhatóak lesznek a dokumentumok egyes részei. Naive Bayes-osztályozással

tesztelve a jellemzőhalmazunk hatékonyságát, 98%-os pontosságot kaptunk egy 100 sorból álló teszthalmaz esetén.

3. Helyesírás-javítás

A dokumentumok alapvető strukturálása és a szöveges tartalmak meghatározása után a következő feladat a dokumentumok normalizálása volt, amelynek első lépése a helyesírási hibák javítása. Esetünkben ez nem csupán a magyar nyelv nehézségeiből eredő problémák megoldására korlátozódott, hanem sok olyan hiba is felmerült a szövegekben, melyek a szakterület sajátosságaiból erednek. A legjellemzőbb hibák az alábbiak voltak:

- elgépelés, félreütés, betűcserék,
- központozás hiányossága (pl mondathatárok jelöletlensége) és rossz használata (pl. betűközök elhagyása az írásjelek körül, illetve a szavak között),
- nyelvtani hibák,
- mondattöredékek,
- a szakkifejezések latin és magyar helyesírással is, de gyakran a kettő valamilyen keverékeként fordulnak elő a szövegekben (pl. *tensio*/*tenzio*/*tensió*/*tenzió*); külön nehézséget jelent, hogy bár egy elvi szabvány létezik ezek helyesírására vonatkozóan, az orvosi szokások változatosak, és még a szakértőknek is problémát jelent az ilyen szavak helyességének megítélése,
- hiányos megfogalmazások gyakori előfordulása, melyek nem tekinthetők a hagyományos értelemben vett rövidítéseknek, azonban teljes szavaknak, kifejezéseknek sem,
- szakterületre jellemző rövidítések, melyeknek sem a jelölés módja, sem a jelentése nem általánosítható.

A fenti hibajelenségek mindegyikére jellemző továbbá, hogy orvosonként, vagy akár a szövegeket lejegyző asszisztensenként is változóak a jellemző hibák. Így elképzelhető olyan helyzet, hogy egy adott szót az egyik dokumentum esetén javítani kell annak hibás volta miatt, egy másik dokumentumban azonban ugyanaz a szóalak egy sajátos rövidítés, melynek értelmezése nem egyezik meg a csupán elírt szó javításával.

A feladat másik nehézségét az jelentette, hogy egyáltalán nem állt rendelkezésünkre nagy méretű helyesen írt klinikai korpusz, ami alapján elő tudtunk volna állítani a javításhoz használható nyelvi és hibamodelleket.

Mivel munkánk jelen fázisában célunk egy kisméretű helyesen írt korpusz előállítása, így a javítási feladatot egy egyszerű lineáris modellel valósítottuk meg. Ehhez különböző nyelvi modelleket kombináltunk, melyeket részben a hibás korpusz alapján építettünk, részben külső erőforrások bevonásával jöttek létre. Az első kettőt a javítás előtti szűrőként alkalmaztuk, a többit pedig a helyes alakok előállításához.

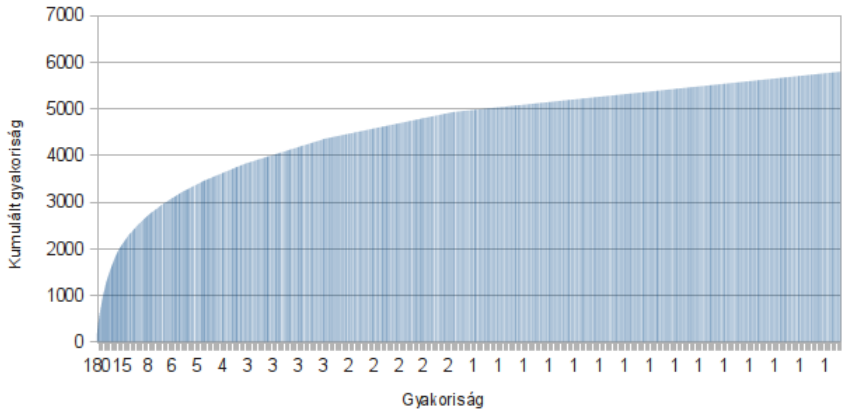
- Stopword lista: az általános stopwordöket kiegészítettük a korpuszra jellemző hasonlóan viselkedő tokenekkel, a leggyakrabban előforduló szóalakok közül kézzel válogatva ki ezeket. Ez elsősorban az írásjel-karaktereket, számokat és egyéb nem szóként vagy rövidítésként kezelendő tokeneket tartalmaz.

- Rövidítéslista: egyszerű mintaillesztéssel kiválasztottuk a potenciális rövidítéseket, majd ezt manuálisan szűrve jött létre a rendszerben használt szóhalmaz. Lehetséges rövidítésnek tekintettük azokat a tokeneket, amik nem mondatvégi szavak, rendelkeznek szó végi ponttal (és esetleg más punktuációval), morfológiai elemző számára ismeretlenek és nem hosszabbak egy előre megadott korlátnál (6 karakter).
- Morfológia által elfogadott szavak listája: kiválogattuk a korpuszból azokat a szóalakokat, amiket a HUMOR morfológiai elemző elfogadott, azaz helyesnek tekinthetőek. Ehhez a morfológiát célszerű volt kiegészítenünk a szakterületre jellemző szavakkal (gyógyszernevek, hatóanyagok, orvosi helyesírási szótár). Az így elfogadott szavak listájából unigram nyelvmodellt építettünk.
- Morfológia által el nem fogadott szavak listája: a fel nem ismert szóalakokból szintén építettünk egy gyakorisági modellt, melyet kétféle módon vettünk figyelembe a javított alakok ajánlása során. Amik kis gyakorisággal fordultak elő ebben a listában, azokat továbbra is rossznak tartottuk, amik azonban nagyon sokszor „rossz" alakban jelennek meg, azokat a morfológiának ellentmondóan, jó alakoknak tekintettük. Így azok a speciális használatú kifejezések, szakszavak, melyeket a morfológia alapján nem ismerünk fel, elfogadottá válhatnak, hiszen a használatuk elég gyakori ahhoz, hogy elfogadottnak tekintsük. A korpuszból generált kumulált előfordulási gyakoriságot reprezentáló görbe gradiensének változása alapján meghatározott küszöbértéknél (2. ábra) nagyobb gyakoriságú szavakat tekintjük helyesnek. A küszöbérték alatti frekvenciájú szavakat pedig $1 - f$ módosított gyakorisággal vettük figyelembe. (Abból a feltételezésből indultunk ki, hogy a legalább n-szer látott tokenek közt fellelhető a szóalakok legnagyobb hányada.)
- Általános és további szakszövegekből álló korpuszok: helyes alakok listájához hasonló gyakorisági modellt építettünk még a Szeged Korpusz alapján, illetve a BNO[3] betegségek listája és leírása alapján is. Itt feltételeztük, hogy csak helyes szóalakokat tartalmaznak.

A modellek létrehozása után a javítandó szöveget egy olyan nyelvfüggetlen tokenizálóval szegmentáltuk, amely képes rövidítések kezelésére a szóalakok és az írásjelek megtartásával egy tokenként, illetve hibatűrő. Érzéketlen a központozási hibákra, hiszen minden nem alfanumerikus karakter mentén – ami nem rövidítés része – új tokent hoz létre. Az fenti eszköz létrehozását az orvosi rekordok különleges nyelvezete (töredékes szerkezetek) és a központozási hibák sűrű megléte indokolta. A szegmentáló egy általános rövidítéslistát és a korábban említett szakterületi rövidítéslistát használja.

A tokenizálás után a stopword-lista és a rövidítéslista alapján kiszűrtük azokat a szavakat, amelyekre nem hajtunk végre javítást. A többi szóalak mindegyikéhez létrejön egy javaslathalmaz, mely az egy Levenshtein távolságra lévő szóalakokat, illetve a morfológia által generált lehetséges javaslatokat rangsorolva tartalmazza. A rangsorolás alapját a fenti modellek és a morfológia által együttesen meghatározott tényező képezi. Mivel minden szóalakra generálunk

[3] Betegségek Nemzetközi Osztályozása



2. ábra. A morfológia által fel nem ismert szóalakok kumulált gyakorisága.

javaslatokat, nem csak azokra, amiket a morfológia rossznak ítél, ezért azt az információt, hogy az eredeti alakot a morfológia elfogadja-e, a javaslatok rangsorolásánál kell figyelembe venni.

A rangsorolás végén a lehetőségek közül az első öt javaslatot tekintettünk lehetséges javításnak. Amennyiben az első és a második helyezett között elég nagy különbség volt, akkor az első javaslatot automatikusan elfogadtuk helyes javításnak, egyébként pedig felhasználói megerősítéssel történt meg a legjobb javaslat kiválasztása az első öt közül.

4. Eredmények

Megvizsgáljuk, hogy a kapott eljárás mint automatikus javító eszköz és mint helyesírási hibákra javaslatot nyújtó eszköz milyen eredményességgel bír. Mivel nem állt rendelkezésünkre helyesen írt szöveg, ezért a kiértékeléshez szükséges teszthalmazt kézzel kellett előállítani. Az eredeti korpusz véletlenszerűen kiválasztott 5%-át javítottuk ki (100 bekezdést). Sok szóalak esetén szembesültünk azzal, hogy gyakran az emberi javítás számára sem egyértelmű, hogy mely alakok fogadhatóak el helyesnek, különösen a vegyes latin–magyar írásmóddal írt szakkifejezéseknél. A módszer eredményeit az általánosan alkalmazott pontosság és fedés alapján értékeltük ki. A pontosság ebben az esetben azt mutatja meg, hogy az első legvalószínűbb javaslatot javításnak tekintve, mekkora a helyesen javított tokenek számának aránya az összes átírt token számához viszonyítva. A fedés értékéből pedig azt tudhatjuk meg, hogy eredeti anyagban lévő hibás tokenek mekkora részét javította a rendszer helyesen. Az F-mérték pedig ezek súlyozott harmonikus közepe. További metrikaként a helyes javaslatok rangját mérve a Mean Average Precision-t (MAP) alkalmaztuk.

1. táblázat. Eredmények az egyes modellek súlyozott kombinációira

OOV	**VOC**	**SZEGED**	**BNO**	**ISORIG**	**HUMOR**	**Pontosság**	**Fedés**	$F_{0.5}$	**MAP**
0,05	0,25	0,15	0,2	0,2	0,15	**0,5555**	**0,8769**	**0,5994**	**0,9863**
0,277	0,277	0	0,166	0,166	0,111	0,5417	0,8769	0,5865	0,9859
0,312	0,312	0	0.187	0.187	0	0.5385	0,8462	0,5807	0,9853

A kiértékelést a lineáris modellünk különböző súlyozott kombinációira vizsgáltuk:

- A morfológiai elemző által elfogadott és nem el fogadott szavak listája (VOC, OOV): Mivel a szövegeinket leginkább az eredeti korpusz jellemzi, ezért az ebből épített modelleket vettük figyelembe a legnagyobb súllyal. A sajátos stílus és szóhasználat miatt mindenképpen a korpuszon belüli előfordulás a hangsúlyosabb az általános szóhasználattal szemben.
- SZEGED, BNO: Mivel a BNO betegségek leírása sok szakkifejezést tartalmaz, viszont sokkal általánosabb formában, mint ahogy az a javítandó szövegekre jellemző, a Szeged Korpusz viszont teljesen általános, hétköznapi kifejezéseket, ezért ezeknek a súlyát kisebb mértékben szükséges figyelembe venni. Az eredményeken látszik, hogy a Szeged Korpusz figyelembevétele valamelyest javít az értékeken, azonban súlyának további növelésével nem érhető el jobb eredmény.
- ISORIG: Az eredetileg feltehetően helyesen írt kifejezések saját maguk valószínűségét erősítik, azonban ennek a tényezőnek a súlyát sem állíthattuk túl nagyra, hiszen ez a morfológia hibáját, illetve szakterületi hiányosságait erősítette volna.
- HUMOR: Jelentősen javított az eredményeken, ha a morfológia által elfogadott javaslatok súlyát megnöveltük. Ehhez szintén a szakkifejezésekkel bővített Humor-t használtuk.

A korpusz sajátos jellegének figyelembevétele miatt - az előzetes feltételezésünknek megfelelően - a meglévő korpuszra épülő modellek(OOV, VOC) magasabb súllyal való figyelembevétele, a morfológiával kiegészítve hozta a legjobb eredményt. (l. 1. táblázat)

A számszerű eredmények nem túl magas értékét több jelenség is magyarázza:

- A teszthalmaz viszonylag kis mérete nem ad teljes képet az összes hibáról, azonban egy nagyobb tesztszöveg létrehozása az emberi erőforrás igénye miatt nehéz.
- A rövidítések felismerésének hiányosságai. Sok esetben nem is értelmezhető a helyesírás-javítás a rövidítések felismerése, a tokenizálás során való helyes kezelése és a feloldás ismerete nélkül. Ilyen mondatok esetén, mint például: *„szemhéjszél idem, mérs. inj. conj, l.sin."* vagy *„Vitr. o.s. (RM) abl. ret. miatt."* a kiértékelés nem tekinthető mérvadónak, azonban a rövidítések megfelelő kezelését a későbbiekben fogjuk megvalósítani.
- Szakterületi többértelműség a latin-magyar vegyes alakok kezelése során. Az *a-á*, *c-k*, *o-ó*, stb. karakterpárok sok esetben egyenértékűek, az ilyen szavaknak sok alakja elfogadott, azonban ez nem fogalmazható meg általános

szabályként. A kiértékelés során minden szónál a gyakrabban előforduló néhány alakját tekintettük helyesnek, ez azonban enyhíthető lenne bármely alak engedélyezésével. Mivel mind az emberi olvasó számára, mind a további alkalmazás céljára alkalmas a jelenlegi módszerrel elérhető valamely forma, így csupán a számértékek növekedése lenne várható ettől, a tényleges minőség javulása nem.

2. táblázat. Példamondatok, automatikus javítással

Hibás mondat	**Automatikusan javított mondat**
A beteg intraorbitalis *implatatumot* is kapott ezért klinikánkon szeptember végén,október elején előzetes *telefonnegbeszélés* után kontrollvizsgálat javasolt.	A beteg intraorbitalis *implantatumot* is kapott ezért klinikánkon szeptember végén,október elején előzetes *telefonmegbeszélés* után kontrollvizsgálat javasolt.
Meibm *mirgy* nyílások helyenként sárgás *kupakszeráűen* elzáródtak, ezeket megint *túvel* megnyitom	Meibm *mirigy* nyílások helyenként sárgás *kupakszerűen* elzáródtak, ezeket megint *tűvel* megnyitom

A javaslatok sorrendjéről elmondható, hogy amikor nem az első eredmény tartalmazza a helyes alakot, akkor az első 5 javaslatban az esetek 99,12%-ban fellelhető a helyes szóalak. Továbbá az információ visszakeresésben használatos MAP metrikával is vizsgálva a találati listánk átlagos pontosságát, a legtöbb esetben 98% fölötti pontosságot kaptunk.

3. táblázat. Automatikus javaslatok hibás szavakhoz

Eredeti szó	**Első javaslat**	**Első öt rangsorolt javaslat**
látahtó	látható	'látható' : 0.1061, 'látahtó' : 0.0004, 'látahető' : 0.0, 'látaptó' : 0.0, 'lgtahtó' : 0.0
rajtra	rajtra	'rajtra' : 0.2631, 'rajta' : 0.1053, 'rajéra' : 0.1052, 'rajtura' : 0.1052, 'rajtja' : 0.10526
implatatumot	implantatumot	'implantatumot' : 0.1053, 'implatatumot' : 0.0009, 'implatatumít' : 0.0, 'őimplatatumot' : 0.0, 'implatáatumot' : 0.0

5. Összefoglalás

A jelenlegi algoritmus célja egy olyan helyesírás-javító alapalgoritmus megvalósítása volt, mellyel egy helyesnek tekinthető orvosi korpusz előállítását tudjuk támogatni. Ezáltal létrehozunk egy olyan szöveget, ami alapján pontosabb hibamodell építhető egy továbbfejlesztett rendszer betanításához.

A javítás egyelőre csupán szószinten történik, a környezet figyelembevétele nélkül. Ahhoz azonban, hogy a környezeteket is fel tudjuk használni az egyes szavak javítása során, egy jó minőségű n-gramokat tartalmazó nyelvmodellre is szükség lenne, aminek előállítása szintén helyes korpuszt igényel.

A javaslatok sorrendjének meghatározásához és azok generálásához, továbbá a modellek felépítéséhez jelenleg csupán teljes szavakat veszünk figyelembe, egy megfelelő hatékonyságú guesser segítségével azonban lemmaszinten is meg lehetne vizsgálni a javaslatok értékét. Ez minden olyan helyzetben segítene, ahol a helyesírási hiba a szótőben fordul elő.

A magyar nyelv agglutináló jellegéből és az összetett szavak írásmódjából adódóan a lehetséges szóalakok kvázi-végtelen száma miatt kézenfekvő volna súlyozott véges állapotú transzducerrel megoldani a javaslatgenerálási feladatot, ami tartalmazná mind a morfológiát, mind az előfordulási gyakoriságokat és a hibamodellt is.

Az elért eredmények alapján bemutattuk, hogy a hosszú távú célként megfogalmazott rendszer kezdeti állapotában is olyan alkalmazásokat tesz lehetővé, amelyek az eredeti dokumentumok kereshetőségében, alkalmazhatóságában, áttekinthetőségében jelentős előrelépést jelentenek. Bemutattuk, hogy egy átfogó, klinikai dokumentumokat elemző rendszer felépítése során a kiindulási állapot létrehozása sem triviális feladat, számtalan nehézséggel kell megküzdeni, ami különösen a kezdeti lépések során mindenképpen igényel emberi munkát is. Az így elérhető egyre nagyobb és egyre pontosabb korpusz javítása azonban fokozatosan teljesen automatikussá válhat.

Hivatkozások

1. Levenshtein, V.: Binary codes capable of correcting spurious insertions and deletions of ones. Problems of Information Transmission **1**(1) (1965) 8–17.
2. Contractor, D., Faruquie, T., Subramaniam, L.: Unsupervised cleansing of noisy text. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics (2010) 189–196
3. Prószéky, G., Novák, A.: Computational Morphologies for Small Uralic Languages. In: Inquiries into Words, Constraints and Contexts., Stanford, California (2005) 150–157.
4. Pirinen, T.A., Lindén, K.: Finite-State Spell-Checking with Weighted Language and Error Models – Building and Evaluating Spell-Checkers with Wikipedia as Corpus. In: Xth SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010. (2010) 13–18.
5. Patrick, J., Sabbagh, M., Jain, S., Zheng, H.: Spelling correction in Clinical Notes with Emphasis on First Suggestion Accuracy. In: 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining. (2010) 2–8.
6. Farkas, R., Szarvas, G.: Automatic construction of rule-based ICD-9-CM coding systems. BMC Bioinformatics **9** (2008)